Report sent on March 7th to: wlcg-scod@cern.ch

Type of Incident: power cut
Location: IN2P3-CC
Duration: 6 minutes (power cut), about 13 hours for total recovery
Date: From February 25th 2011, 23h10 to February 26th 2011, 12h25
Author: Rolf Rumler

Description

Due to a temporary overload a fuse cut off the power supply for air conditioning and cooling devices. This induced an emergency shutdown for most of the batch worker nodes.

Timeline

February 25th (Friday)

- 23:10 A fuse cut off the power supply for air conditioning and cooling devices.
- 23:16 The quickly rising temperature triggered an emergency shutdown for most of the batch system's worker nodes. Only 15 percent of them stayed available but all other services continued without problems, especially data transfer and storage.
- 23:16 The fuse got re-armed, air conditioning restarted.

February 26th (Saturday)

- 00:05 First message to staff from engineer on duty.
- 01:25 Message to registered users about the incident.
- From about 09:00 onward, power on and verification of worker nodes.
- 12:25 All worker nodes are back. Still problems on other machines, not foreseen for usage by WLCG.

Analysis

Normal usage of a power outlet caused an overload of a fuse already charged by cooling devices. Stopping the use of the outlet and than re-enabling the fuse was sufficient to restart the air condition. However in the meantime the temperature in the machine room raised to critical levels. To avoid damage, especially to storage devices, the automatic emergency shutdown had been set to react at the temperature level reached; most of the worker nodes were powered off except the remotely hosted ones (at the CINES centre of Montpellier).

Simply powering them on again would not have been sufficient as by experience various machines would not come up correctly, so that every machine needed verification. This delayed the largely manual restart to the morning hours of Saturday.

Communication about the incident was correct with respect to the registered users of the IN2P3-CC but not sufficient for the grid-only users.

Impact

The compute service provided by the worker nodes is the less critical one of the tier1 because most of the jobs crashing when the workers are stopped can be restarted, either locally or elsewhere on the grid. All CEs continued to work as every other service, too, so only the CPU time of the jobs running at the time of the incident was lost.

The CPU capacity of the tier1 stayed at about 15 percent for about 12 hours.

Corrective actions

It will be checked whether a better separation of critical power supply paths from ordinary ones can be implemented rapidly.

The outage had not been declared as a downtime. This was due to an inappropriate documentation which did not state clearly that worker node failures had to be associated to CEs and a downtime announced for them, even if the compute elements themselves were not impacted by the incident.